Chemical Science and Technology Laboratory
National Institute of Standards and Technology

Chemical and Allied Products
Data and Informatics

The IUPAC International Chemical Identifier (InChI)

NIST, in collaboration with International Union of Pure and Applied Chemistry (IUPAC), has created a chemical identifier standard that could be adopted by the entire chemical community known as the IUPAC-International Chemical Identifier (InChI). The goal of this work was to create a naming system that would allow computers to uniquely identify a chemical, based entirely on the connectivity of the molecule, and independent of how it is drawn.

D. Tchekhovskoi (Div. 838)

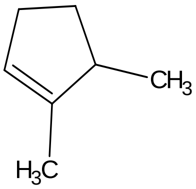
The question of clearly identifying a chemical has been present almost since the beginnings of modern chemistry. As the number of chemicals grew, the need for systematic naming produced a number of results. The most widely adopted of these is that of the International Union of Pure and Applied Chemistry (IUPAC). But for many chemicals, the resulting names are complex and so common names are still widely used. For most chemists, the graphical structure is the best method for identifying a chemical. The structure provides graphic information that can allow a rapid understanding of the properties of the chemical that a long text name can never provide. However, the phenomena of tautomerism and (de)protonation change chemical structure, thus hiding the chemical identity of compounds. In addition, the same chemical structure may be drawn in such different ways that it is hard to establish equivalence of the drawings.

The need for a uniform and open standard that could be adopted by the entire chemical community generated the NIST/IUPAC project to develop a chemical identifier – The IUPAC-International Chemical Identifier (InChI). The aim of the project was not to create another naming system, or at least not a naming system that would be usable in normal communication. The goal was to create a naming system that would allow computers to uniquely identify a chemical, regardless of how it is drawn based entirely on the connectivity of the molecule – that is what atoms are connected to what other atoms.

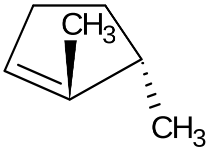
In the InChI software, much of what is normally viewed as "chemical information" is discarded, and the molecules are trimmed to the minimum information needed to differentiate one from the other. A layered approach deals with some of the more complex issues of chemical structure.

For example, the two molecules in the figure differ only in that one has the two methyl groups on the same side relative to the plane of the ring (on the right – called *cis*) and the other has the two methyl groups on opposite sides of the ring (in the center – called *trans*). On the left, is a diagram that can be used to represent either of the molecules.

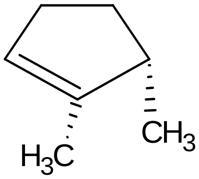
The left diagram shows only the connectivity and does not specify if the molecule is *cis* or *trans*. The problem encountered prior to InChI is that the data retrieval was often dependent upon the way the molecule had been drawn. There is often a need to distinguish between the *cis* and *trans* form, and often a need to search for all possible forms, including cases where the configuration of the molecule was not known or it was known that a mixture was present.



unspecified



trans



cis

The approach taken in developing InChI is a layered approach. This allowed as much information as was known to be specified, the search could be performed only on the information known and the search could be stopped with less than full information. Thus in the case above, a search for the *cis* isomer could be allowed to stop when it matched the connectivity or continued to find only the molecules that matched the geometric isomer.

The IUPAC International Chemical Identifier has been released by IUPAC in April 2005.

An InChI validation protocol was released in August 2006. It was designed to verify that a 3rd party InChI software or a ported to other platforms InChI software would still generate valid InChI strings. Simultaneously was released a minor software update. It includes a facility that allows reconstructing of a chemical structure out of InChI.

The IUPAC InChI has been adopted/used by a variety of entities:
PubChem, a major resource for medical research at the National Institues of Health (NIH) has adopted InChI as a standard for identifying and searching for compounds: For example,see:
http://pubchem.ncbi.nlm.nih.gov/summary/summary.cgi?cid=6986.

It has been integrated by ACD Labs in their widely used commercial drawing program, ChemSketch as well as the freeware version that is distribute for home and student use, see: http://www.acdlabs.com/download/chemsk.html.
The Chemistry WebBook had adopted InChI and displays the identifier for all data in the WebBook thus making it possible for the very large audience that uses the WebBook to gain access to the identifier, see:
http://webbook.nist.gov/cgi/cbook.cgi?Name=Horse&Units=SI,
The Environmental Protection Agency has adopted InChI for use in their Distributed Structure-Searchable Toxicity (DSSTox) Public Database Network, see:
http://www.epa.gov/nheerl/dsstox/MoreonINChI.html
The Web of Science, (Thomson Scientific) one of the most widely used information sources for searching the scientific literature has adopted the identifier.

There are many other uses of the identifier – and since the software is distributed free of charge we do not even know if we are aware of all of them. One example is the Compendium of Pesticide Common Names (Alan Wood, UK), a site that allows users to find the common name as well as the structure for a pesticide, see:
http://www.hclrss.demon.co.uk/. The wide number of adoptions has been made easy by a range of software solutions developed as a part of the project. Many of these are available from IUPAC.

Impact:
The identifier project is an example of open source software. Since it is a standard, it is not wise or reasonable to modify the code, but the code is available for any who want to use it. In addition, a number of tools to allow users to easily analyze their own structures using the freely available compiled code. As InChI becomes more widely adopted, it is expected that it will enable a standardized referencing and search for chemical structures both over the Internet and in proprietary databases.

.Future Plans:
We are examining the best way to extend InChI to include polymers, phase, and excited states. In addition, we are examining ways to make separate parts of InChI algorithm – chemical structure normalization, canonicalization, and serialization – available to 3rd party software through InChI API (Application Programming Interface.)

Project Team: D. Tchekhovskoi, S. Stein, S. Heller (838)

Publications and Presentations:

- The IUPAC International Chemical Identifier: InChl — A New Standard for Molecular Informatics, Alan McNaught, *Chemistry International*, November-December 2006, **8**(6), 12-14.
- Using InChI, Jeremy Frey, *Chemistry International*, November-December 2006, **8**(6), 14-15.
- Chemical 'Naming' Method Unveiled, *Chem. & Eng. News*, 22 Aug 2005 [link to *C&EN*]
- Analysis of a Set of 2.6 Million Unique Compounds gathered from the Libraries of 32 Chemical Providers, A. Monge, A. Arrault, C. Marot and L. Morin-Allory, presented at the *10th Electronic Computational Chemistry Conference*, April 2005
- International chemical identifier goes online, *Chem. World*, 16 May 2005 [link to *CW*]
- Application of InChI to Curate, Index, and Query 3-D Structures, M.D. Prasanna, J. Vondrasek, A. Wlodawer and T.N. Bhat, *Proteins: Structure, Function, and Bioinformatics*, 2005, **60**, 1-4
- Enhancement of the chemical semantic web through the use of InChI identifiers, S.J. Coles, N.E. Day, P. Murray-Rust, H.S. Rzepa and Y. Zhang, *Org. Biomol. Chem.*, 2005, **3**(10), 1832-1834
- InChI FAQ, by Nick Day (Unilever Centre for Molecular Informatics, Cambridge University)
- Representation and Use of Chemistry in the Global Electronic Age, P. Murray-Rust, H.S. Rzepa, S.M. Tyrrell and Y. Zhang, *Org. Biomol. Chem.*, 2004, 3192-3203 [www.ch.ic.ac.uk/rzepa/obc/]
- That InChI feeling, *Reactive Reports*, issue 40, Sep 2004
- Unique labels for compounds, *Chem. & Eng. News*, 2 Dec 2002
- Chemists synthesize a single naming system, *Nature*, 23 May 2002
- That InChI feeling ... *The Alchemist*, 24 Apr 2002
- What's in a Name? *The Alchemist*, 21 Mar 2002